A REPORT OF ONE MONTH TRAINING

at

EXCELLENCE TECHNOLOGY, MOHALI

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF

BACHELOR OF TECHNOLOGY

(Computer Science & Engineering)

JUNE-JULY 2025

SUBMITTED BY:

Lovepreet Singh

URN: 2302598

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

GURU NANAK DEV ENGINEERING COLLEGE, LUDHIANA

(An Autonomous College Under UGC ACT)

Certificates By Excellence Technology

EXCELLENCE TECHNOLOGY

NATIONAL AWARD WINNER

Ref. no. Extech/1308/7299

CERTIFICATE
OF TRAINING

PROUDLY PRESENTED TO

Lovepreet Singh

S/o/D/o Sh. Beant Singh

of Guru Nanak Dev Engg. College, Ldh. has successfully completed his/her training on Data Analysis from 23.06.25 to 07.08.25 during the tenure of the above training, we have found him/her a hardworking

Ministry of MSME, Govt. of India
Reg. no. - CH01D0007051

ISO 9001:2015

Director

Grade - A☐ A+☑ B☐ B+☐ C☐

info@excellencetechnology.in www.excellencetechnology.in


CERTIFICATE
OF COMPLETION
Ref. no. Extech/1308/ 7299
EXCELLENCE TECHNOLOGY
IAF
eiaCI
PROUDLY PRESENTED TO
LOVEPREET SINGH
S/o/D/o Sh. BEANT SINGH
of GURU NANAK DEV ENGG. COLLEGE has successfully
completed his/her One Hundred Twenty Hours (120 hrs.) ISO certificate course
DATA ANALYSIS
from 23.06.25 to 07.08.25 during the tenure
of the above Course, we have found him/her a hardworking
★ISO★
9001 - 2015
CERTIFIED
Training Co-ordinator
Director
9001:2015
info@excellencetechnology.in
www.excellencetechnology.in

*


CERTIFICATE
OF PERSONALITY DEVLOPMENT
EXCELLENCE TECHNOLOGY
Ref. no. Extech/1308/ 7299
PROUDLY PRESENTED TO
LOVEPREET SINGH
Has Successfully Completed Personality Development Course.
from 23.06.25 to 07.08.25
PERSONAL GROWTH
SIGNATURE
info@excellencetechnology.in
www.excellencetechnology.in

GURU NANAK DEV ENGINEERING COLLEGE, LUDHIANA
CANDIDATE'S DECLARATION

I "LOVEPREET SINGH " hereby declare that I have undertaken one month training "EXCELLENCE TECHNOLOGY" during a period from 26-06-2025 to 07-08-2025 in partial fulfilment of requirements for the award of degree of B.Tech (Computer Science & Engineering) at GURU NANAK DEV ENGINEERING COLLEGE, LUDHIANA. The work which is being presented in the training report submitted to Department of Computer Science & Engineering at GURU NANAK DEV ENGINEERING COLLEGE, LUDHIANA is an authentic record of training work.

Signature of the Student

The one month industrial training Viva–Voce Examination of_____has been held on _____and accepted.

Signature of Internal Examiner

Signature of External Examiner

CONTENTS

ABSTRACT

The industrial training undertaken during this period provided an in-depth practical exposure to the field of Data Analytics and its real-world applications. The primary focus of the training was to acquire knowledge of data handling, analysis, visualization, and interpretation using modern analytical tools such as Python, Pandas, NumPy, Matplotlib, and Seaborn. The project component of the training centered around performing Exploratory Data Analysis (EDA) on the widely recognized Titanic Dataset, which served as a comprehensive case study for understanding the end-to-end data analysis workflow.

The project began with understanding the structure and nature of the dataset, followed by thorough preprocessing steps including data cleaning, handling missing values, and analyzing categorical and numerical variables. A detailed univariate and bivariate analysis was carried out to discover relationships and correlations among key features such as age, gender, passenger class, and survival status. Further, feature engineering techniques were applied to create new variables like Family Size and Fare per Person, which added depth to the analysis and enhanced interpretability.

Through this project, meaningful insights were drawn about survival patterns on the Titanic, demonstrating how data-driven techniques can extract valuable information from raw datasets. The training not only strengthened technical competence in data analytics but also developed critical thinking, problem-solving, and data storytelling abilities essential for modern analytical roles. This experience bridged the gap between theoretical learning and industrial application, preparing the trainee for future challenges in the domain of Data Science and Artificial Intelligence.

ACKNOWLEDGEMENT

I take this opportunity to express my sincere gratitude to all those who have supported and guided me throughout my 45 days of industrial training at Excellence Technology, Mohali. This training has been an enriching experience that has significantly enhanced my technical knowledge and understanding of Data Analytics and its real-world applications.

First and foremost, I would like to extend my heartfelt thanks to the faculty members and training coordinators of Excellence Technology for providing me with this valuable learning opportunity. Their continuous guidance, expert supervision, and constructive feedback throughout the training period were instrumental in helping me grasp the core concepts of data analysis, visualization, and interpretation.

I am deeply grateful to my mentor and instructors at Excellence Technology for their constant encouragement, patience, and willingness to share their expertise. Their insightful explanations and practical demonstrations helped me build a strong foundation in Python, Pandas, NumPy, Matplotlib, and Seaborn, as well as understand the systematic workflow of Exploratory Data Analysis (EDA).

I would also like to express my appreciation to my college authorities and faculty members for motivating me to undertake this industrial training and for their continuous support and cooperation. Their emphasis on experiential learning has been invaluable in shaping my technical and analytical skills.

Finally, I extend my deepest gratitude to my family and friends for their encouragement, understanding, and moral support during the entire training period. Their motivation helped me stay focused and complete my project successfully.

ABOUT THE COMPANY

Excellence Technology is a leading industrial training and software development company located in Mohali, Punjab, known for providing high-quality professional training and technical education in emerging fields such as Data Analytics, Artificial Intelligence, Machine Learning, Web Development, Python Programming, Android Development, Digital Marketing, and various other cutting-edge technologies.

Established with the vision of bridging the gap between academic knowledge and industry requirements, Excellence Technology focuses on imparting practical, project-based learning to students, fresh graduates, and working professionals. The company's training modules are carefully designed by experienced industry experts to align with current technological trends and corporate expectations.

The organization offers hands-on industrial training programs, internships, and career-oriented workshops that equip learners with the skills required to meet real-world challenges. The training approach combines theoretical understanding with practical implementation through live projects, case studies, and collaborative exercises.

Excellence Technology has successfully trained thousands of students across India, helping them secure placements in reputed IT firms and startups. The institute also emphasizes soft skills development, including communication, teamwork, and professional ethics, to ensure holistic personality growth of its trainees.

The company's infrastructure includes modern computer labs, high-speed internet facilities, and expert trainers who provide personalized guidance to each learner. With a mission to promote skill-based learning and enhance employability, Excellence Technology continues to be one of the most trusted names in professional and industrial training in Northern India.

LIST OF FIGURES

LIST OF TABLES

DEFINITIONS, ACRONYMS AND ABBREVIATIONS

Term / Acronym	Full Form / Definition
EDA	Exploratory Data Analysis – The process of analyzing datasets to summarize their main characteristics through visualization and statistics.
CSV	Comma-Separated Values – A text file format used to store tabular data where each line represents a record.
ML	Machine Learning – A subset of artificial intelligence enabling systems to learn and improve from data without explicit programming.
Pandas	A Python library used for efficient data manipulation and analysis.
NumPy	A Python library that supports large, multi-dimensional arrays and matrices with mathematical functions.
Matplotlib	A Python plotting library used to create static and interactive visualizations.
Seaborn	A data visualization library based on Matplotlib that provides advanced and aesthetically pleasing statistical plots.
Feature Engineering	The process of creating new features from existing data to improve model performance and insights.
Bivariate Analysis	A statistical method used to study the relationship between two variables.
Titanic Dataset	A publicly available dataset from Kaggle containing passenger details from the RMS Titanic, commonly used for data analysis and prediction exercises.

CHAPTER 1

INTRODUCTION

1.1 Introduction to Data Analytics Training

The rapid advancement of technology and the exponential growth of data in every sector have given rise to a new discipline known as Data Analytics. It represents the convergence of mathematics, statistics, programming, and business intelligence — all directed toward one fundamental goal: extracting meaningful insights from raw data to support decision-making. Organizations today generate massive amounts of information through digital transactions, customer interactions, and online activities. Data analytics enables them to utilize this vast data effectively to improve performance, optimize operations, and gain competitive advantages.

The Data Analytics training undertaken during the industrial program at *Excellence Technology, Mohali* aimed to equip trainees with both the conceptual foundation and practical experience required to work with real-world datasets. The training emphasized understanding the complete data analytics lifecycle — from data collection to visualization — while developing problem-solving, analytical, and technical skills essential for today's data-driven industries.

During the training, students were introduced to the importance of data as the new form of organizational capital. Companies across domains such as finance, healthcare, e-commerce, manufacturing, and education rely heavily on data analytics to understand customer behavior, detect trends, forecast outcomes, and make evidence-based decisions. Through a well-structured curriculum, the training helped participants comprehend how analytical thinking transforms raw, unorganized information into actionable insights.

The course began with an introduction to the fundamentals of data analytics, explaining various types of data — numerical, categorical, and textual — and the importance of data preprocessing. Students learned about the challenges associated with raw data, such as missing values, duplicates, and inconsistencies, and how to clean and prepare data for analysis. Emphasis was

placed on the fact that nearly 70–80% of the analytical process involves data cleaning and preparation, making it one of the most critical stages in any analytics project.

Once the basics were established, participants were trained to explore datasets using exploratory data analysis (EDA) techniques. EDA helps in understanding the structure, distribution, and relationships between different variables. By using visualization tools, learners could identify hidden trends, outliers, and correlations within datasets. The training encouraged an inquisitive approach — asking the right questions about data before attempting to build models or draw conclusions. This stage built the analytical mindset necessary for interpreting real-world data problems effectively.

A key focus of the training was on the use of modern analytical tools and software that are standard in the industry. The participants worked extensively with Python and its libraries such as Pandas, NumPy, Matplotlib, and Seaborn, which form the backbone of most data analytics workflows. Python's versatility allowed trainees to perform data manipulation, statistical analysis, and visualization seamlessly within a single environment.

Alongside Python, SQL/MySQL was used to handle structured databases, query large tables efficiently, and integrate multiple data sources. Students also learned to perform preliminary analysis in Microsoft Excel, utilizing pivot tables, charts, and formulas for summary statistics. To develop visualization and reporting skills, tools such as Power BI and Tableau were introduced, enabling participants to create professional dashboards and communicate analytical results effectively to non-technical audiences.

Another important aspect of the training was feature engineering and transformation, where raw data is converted into a more informative format suitable for analysis or modeling. Trainees practiced deriving new features, normalizing data, and encoding categorical variables. Through hands-on sessions, they understood how these preprocessing steps directly influence the quality of insights obtained from the data.

Collaboration was an integral part of the training experience. Students worked in teams to discuss analytical challenges, brainstorm possible solutions, and validate findings. This collaborative environment mirrored the industrial setting where cross-functional teamwork is essential for problem-solving. Regular feedback sessions and mentor guidance provided clarity and direction, helping participants to refine their analytical approach and improve technical accuracy.

Throughout the 45-day period, the training followed a progressive learning structure. Each week focused on a different stage of the analytical workflow — starting from data understanding, followed by data cleaning, analysis, visualization, and finally report generation. By the end of the program, participants were capable of independently performing an end-to-end data analysis project. The final stage involved applying all acquired knowledge in a practical project titled "Exploratory Data Analysis on the Titanic Dataset." This project acted as a capstone experience, combining theoretical concepts, data handling techniques, and visualization skills into one cohesive analysis.

Beyond technical training, the course also emphasized interpretation and communication of analytical results. Students learned that analytics is not only about performing calculations but about telling a story through data. The ability to convey findings clearly and persuasively — using visuals, summaries, and insights — is what differentiates a good analyst from a mere technician. The training encouraged the habit of documenting observations, creating logical narratives, and making data-driven recommendations.

Overall, the Data Analytics training was a comprehensive program designed to build both competence and confidence. It helped participants appreciate the real-world relevance of analytics and its impact across industries. By the end of the training, students were proficient in working with structured and unstructured datasets, visualizing relationships between variables, and deriving insights that could guide decision-making.

Most importantly, the training cultivated a data-driven mindset — an analytical way of thinking that values evidence over assumption and reasoning over intuition. This mindset will serve as a foundation for future learning and professional growth in the field of data science and analytics.

1.2 Background of Data Analytics

In today's digital era, data has emerged as the most valuable resource. Organizations generate massive volumes of data daily through business transactions, customer interactions, supply chain operations, healthcare systems, and numerous other sources. However, raw data in itself does not hold value unless it is processed, analyzed, and converted into meaningful insights. This is precisely where Data Analytics plays a significant role.

Data Analytics refers to the systematic computational analysis of data to identify meaningful patterns, trends, and correlations. It involves applying statistical techniques, algorithms, and modern computational tools to extract actionable information from raw data. Businesses utilize analytics to make informed decisions, optimize performance, predict future trends, and maintain competitiveness.

Historically, analytics began with manual calculations and simple tabulations. Over time, with the rise of computers and business intelligence tools, analytics became more structured and sophisticated. Today, the availability of large-scale data, coupled with advancements in machine learning and artificial intelligence, has transformed data analytics into a cornerstone of digital innovation.

The importance of data analytics can be seen across multiple domains:

- **Healthcare**: Predicting disease outbreaks, analyzing patient records, recommending treatments.
- **Finance**: Fraud detection, risk assessment, portfolio optimization.

- **Retail and E-commerce**: Customer segmentation, personalized recommendations, inventory optimization.
- **Transportation**: Route optimization, predictive maintenance, demand forecasting.
- **Social Media and Marketing**: Sentiment analysis, campaign effectiveness measurement.

Thus, the scope of data analytics is immense, making it a highly relevant and in-demand skill in the present and future job markets.

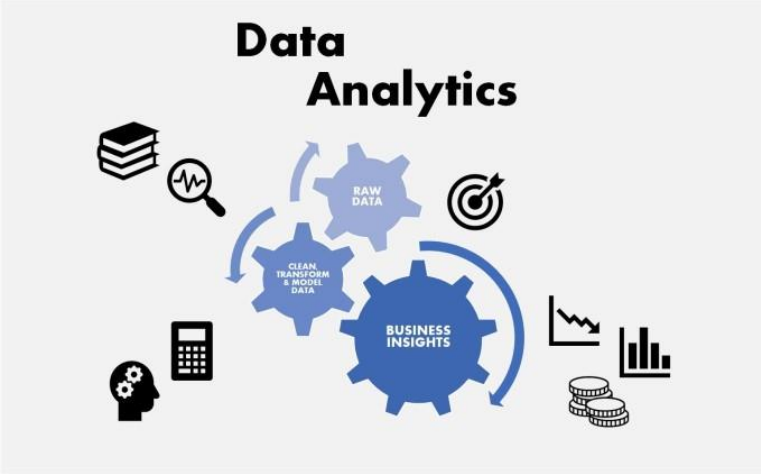


Figure 1.1: Transformation of Data into Insights through Analytics

1.3 Importance of Data Analytics in Industry

In modern enterprises, data is often called the "new oil" because of its transformative potential. With proper analysis, organizations can turn raw data into a strategic asset that drives innovation and competitiveness. Some of the key roles of data analytics in industries include:

1. **Data-driven Decision Making**: Instead of relying solely on intuition, managers and policymakers now use analytics-driven dashboards and reports to take evidence-based decisions.
2. **Trend Forecasting**: Predictive analytics enables businesses to forecast sales, customer demand, and market fluctuations.

3. **Operational Efficiency**: By identifying inefficiencies, companies can reduce costs, improve productivity, and optimize resource allocation.
4. **Customer Insights**: Analytics reveals customer preferences, buying behaviors, and satisfaction levels, which helps in designing better products and services.
5. **Risk Management**: In sectors like banking and insurance, analytics helps in fraud detection, risk assessment, and compliance monitoring.

The above applications highlight that analytics is not limited to IT companies but is equally important in every industry where data is generated. Therefore, undergoing industrial training in this field provides students with a competitive advantage in their career.

1.4 Objectives of the Training

The main purpose of this 45-day training program was to gain theoretical knowledge and practical exposure in the field of data analytics. The objectives of the training were:

- To understand the concepts and processes of data analytics.
- To learn the different types of analytics: Descriptive, Diagnostic, Predictive, and Prescriptive.
- To acquire hands-on skills in widely used analytics tools such as Python, SQL, and Excel.
- To practice data preprocessing, cleaning, and visualization techniques on real datasets.

1.5 Theoretical Explanation of Data Analytics Concepts

Data Analytics is not a single-step activity; rather, it is a structured process that involves multiple stages. The major components include:

1. **Data Collection** – Gathering raw data from databases, sensors, online platforms, or business transactions.

2. **Data Cleaning and Preprocessing** – Handling missing values, removing duplicates, normalizing formats, and ensuring quality.
3. **Data Transformation** – Converting data into meaningful structures (tables, charts, graphs) for analysis.
4. **Data Analysis** – Applying statistical models, algorithms, and programming techniques to identify insights.
5. **Data Visualization and Reporting** – Representing results in the form of dashboards, graphs, and summaries for decision-making.

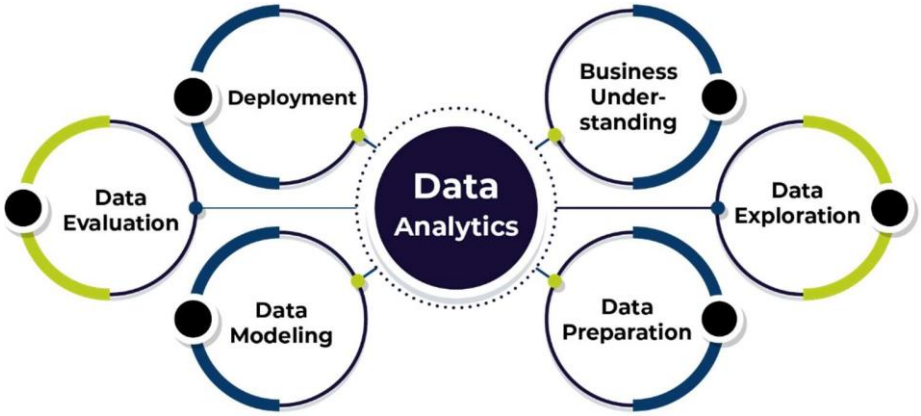


Figure 1.2 Components of Data Analytics

1.5.1 Types of Data Analytics:

- **Descriptive Analytics**: Explains "what happened" in the past.
- **Diagnostic Analytics**: Explains "why it happened."
- **Predictive Analytics**: Predicts "what is likely to happen."
- **Prescriptive Analytics**: Suggests "what should be done."

This theoretical foundation formed the basis for the practical training modules undertaken at Excellence Technology.

Table 1.1: Types of Data Analytics

Type of Analytics	Purpose	Example
Descriptive	Explains what happened	Monthly sales reports
Diagnostic	Explains why it happened	Drop in sales due to high prices
Predictive	Predicts what will happen	Forecasting next quarter's revenue
Prescriptive	Suggests what should be done	Recommending promotional strategies

1.6 Software Tools Learned

During the industrial training at Excellence Technology, Mohali, several essential tools and software platforms were learned and utilized. Each played a vital role in understanding, analyzing, and visualizing data effectively. The combination of these technologies provided a strong foundation in data analytics, enabling efficient data handling, visualization, and interpretation.

The following tools were integral to the training process:

1.6.1 Python Programming Language

Python is one of the most widely used programming languages in data analytics and machine learning. Its simplicity, versatility, and vast library support make it an ideal choice for performing end-to-end data analysis tasks.

During the training, Python was used extensively for data cleaning, transformation, visualization, and basic statistical analysis. Several popular libraries were explored:

- **NumPy (Numerical Python):**

  NumPy provides efficient array and matrix operations. It supports mathematical and

logical operations on large datasets and is the foundation for most data manipulation tasks in Python. It was used for operations such as handling missing data, performing aggregations, and conducting mathematical computations.

- **Pandas (Python Data Analysis Library):**

  Pandas was the core library used for importing, cleaning, and transforming datasets. It provides high-level data structures like DataFrames, which simplify data manipulation tasks such as filtering, grouping, merging, and reshaping. The Titanic dataset project heavily relied on Pandas for exploratory data analysis and preprocessing.

- **Matplotlib:**

  Matplotlib is a powerful data visualization library used to create a variety of static, animated, and interactive plots. It was employed to draw bar charts, histograms, and line graphs, which helped to visualize relationships and trends in the dataset clearly.

- **Seaborn:**

  Seaborn is built on top of Matplotlib and provides a more attractive, high-level interface for drawing informative and aesthetic statistical graphics. It was primarily used to create correlation heatmaps, violin plots, and boxplots for the Titanic dataset. Its integration with Pandas made data visualization more convenient and insightful.

Together, these Python libraries enabled comprehensive data analysis — from data cleaning and manipulation to visualization and interpretation.

1.6.2 Microsoft Excel

Excel remains one of the most popular tools for preliminary data analysis, especially in the corporate and business environments. During the training, Excel was used for performing initial data exploration, descriptive statistics, and visualization before moving to advanced Python- based analytics.

Key activities included:

- Using Pivot Tables to summarize and organize data.
- Applying conditional formatting to highlight trends and anomalies.
- Using formulas and functions such as AVERAGE, COUNTIF, VLOOKUP, etc.
- Creating charts and graphs (like bar charts and pie charts) for data presentation.

Excel provided a clear understanding of how data can be summarized and visualized manually before automating these tasks through programming. This bridge between traditional and modern analytical approaches proved essential for a well-rounded understanding of data analytics.

1.6.3 Jupyter Notebook

The entire training and project execution were carried out using Jupyter Notebook, an open- source web application that allows users to create and share documents containing live code, equations, visualizations, and explanations. It was accessed through the Anaconda distribution, which provides a user-friendly environment for managing Python packages and dependencies.

Jupyter Notebook was chosen for:

- Writing and executing Python code interactively.
- Documenting the analysis process with markdown and comments.
- Visualizing the results within the same environment.

Anaconda, on the other hand, simplified environment management by pre-installing essential libraries like NumPy, Pandas, and Matplotlib. This integrated setup helped in maintaining a smooth workflow throughout the training.

1.6.4 Power BI / Tableau

To enhance the visualization and presentation of analytical results, Power BI and Tableau were introduced during the training as professional business intelligence tools. These tools are used for dashboard creation, data visualization, and interactive reporting.

Key learnings included:

- Connecting datasets from Excel, CSV, or SQL databases.
- Cleaning and transforming data using built-in tools.
- Designing interactive dashboards with filters, slicers, and dynamic visuals.
- Using charts, maps, and KPI indicators to summarize data effectively.

These tools provided hands-on experience in presenting analytical results in a visually engaging and professional manner, which is crucial for decision-making in real-world business scenarios. Power BI and Tableau bridged the gap between technical analysis and business reporting, demonstrating how data analytics supports actionable insights.

CHAPTER 2

TRAINING WORK UNDERTAKEN

2.1 Overview of Training Modules

The 45-day industrial training program at Excellence Technology, Mohali was systematically designed to provide students with comprehensive, hands-on exposure to the domain of Data Analytics. The training structure followed a progressive learning model, beginning with fundamental programming concepts and gradually advancing toward complex analytical techniques. Each week introduced a new module that built upon the concepts learned in the previous one, ensuring a smooth and structured learning curve for all participants.

The course commenced with the foundations of Python programming, focusing on the syntax, data types, and core logic required for analytical problem-solving. Once participants developed confidence in programming, the training moved toward data handling, cleaning, preprocessing, and visualization techniques, which form the backbone of any data analytics workflow. Later modules emphasized database management using SQL and concluded with statistical analysis, enabling participants to extract meaningful insights from datasets.

The overall training methodology was highly practical-oriented, emphasizing implementation and experimentation rather than rote theoretical learning. Each concept introduced in the sessions was accompanied by live demonstrations, coding exercises, and real-world data challenges. Students were given daily practice assignments to reinforce classroom learning, while weekly projects helped them integrate multiple concepts into coherent analytical workflows. Trainers maintained an interactive approach, encouraging students to ask questions, share ideas, and collaborate effectively on tasks.

To ensure a well-rounded learning experience, doubt-clearing sessions and peer discussions were conducted regularly, helping students overcome challenges encountered during their practice. Regular progress evaluations and feedback from trainers allowed participants to identify their

strengths and areas of improvement, ultimately enhancing both their technical and analytical capabilities. These feedback mechanisms also simulated a professional work environment, where continuous improvement and adaptability are essential for success.

In addition to the technical modules, students were also introduced to the real-world applications of Data Analytics across various industries, including business, finance, e-commerce, and healthcare. Trainers frequently cited case studies to illustrate how data-driven decision-making is transforming organizations globally. This contextual understanding helped learners connect theoretical knowledge with practical outcomes, making the training experience both meaningful and industry-relevant.

The training environment at Excellence Technology was highly engaging, collaborative, and dynamic. Students worked on multiple datasets, exploring data structures, performing statistical operations, and visualizing insights through Python libraries like Pandas, NumPy, Matplotlib, and Seaborn. Each week concluded with review sessions where students showcased their work, presented their findings, and discussed challenges faced during implementation. This practice not only improved their technical presentation skills but also instilled confidence in articulating analytical reasoning effectively.

Moreover, the training emphasized professional discipline and workflow management, such as maintaining clean code, documenting each analytical step, and adhering to deadlines. Trainers encouraged students to adopt best practices used in the analytics industry, including version control, modular programming, and reproducible analysis. These elements helped bridge the gap between academic learning and industrial expectations, preparing students for real-world roles in data-driven environments.

By the end of the 45-day period, participants had developed a strong foundational understanding of data analytics, covering every essential phase — from data acquisition and cleaning to visualization and interpretation. They were not only able to handle large datasets independently

but also to draw meaningful conclusions and insights from them. The program successfully instilled analytical thinking, problem-solving skills, and professional confidence among trainees, laying a solid foundation for future careers in data analytics and related domains.

The following table provides a detailed overview of the week-wise training modules, highlighting the major activities, focus areas, and tools/technologies introduced during the course.

Table 2.1 Week-wise Training Module

Week	Module	Key Activities	Tools /Techniques
1	Python Basics	Variables, data types, operators, lists, tuples, dictionaries	Python, Jupyter Notebook
2	Pandas & NumPy	DataFrame creation, CSV reading, basic data operations, array computations	Pandas, NumPy
3	Data Cleaning & Preprocessing	Handling missing data, duplicates, outliers, encoding categorical data	Pandas
4	Data Visualization	Bar plots, histograms, scatter plots, box plots, heatmaps	Matplotlib, Seaborn
5	SQL & Database Management	CRUD operations, SELECT queries, joins, aggregations	MySQL

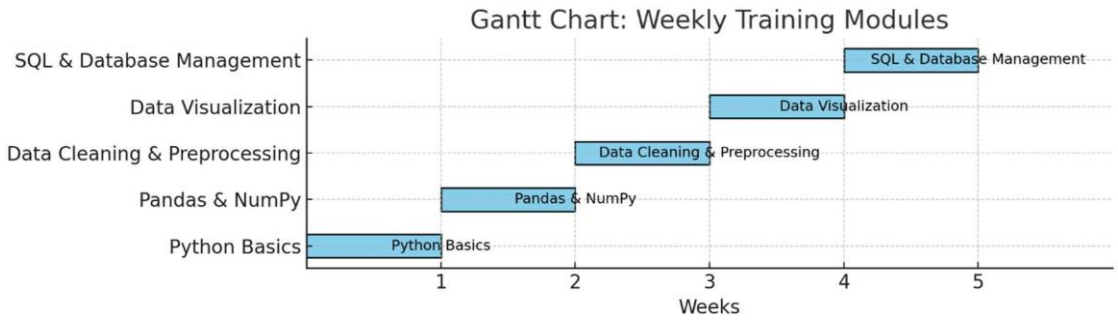


Figure 2.1 Weekly Training Module

2.2 Python Programming Module

Objective: Develop programming skills for data analytics.

The training began with a comprehensive introduction to the Python programming language, which is one of the most popular tools used in the field of Data Science and Analytics. Participants learned about Python syntax, indentation, data types (integers, floats, strings), and various operators used for computation and logic building.The concept of data structures was an important component of this module. Students worked with lists, tuples, sets, and dictionaries— understanding their properties, use-cases, and efficiency in handling data. The module also covered control statements such as if-else, for, and while loops, along with user-defined functions to modularize code.Trainers emphasized file handling and basic data input/output operations, especially reading and writing CSV files. This laid the groundwork for later modules that involved data manipulation using Pandas.

Activities Undertaken:

- Learning Python syntax and operators.
- Working with data structures: lists, tuples, dictionaries, and sets.
- Implementing control structures: if-else, loops, functions.
- Reading, writing, and manipulating CSV files using Pandas.

Learning Outcome:

- Ability to read, manipulate, and inspect datasets.
- Gained foundational skills in programming logic and Python syntax.

2.3 Pandas & NumPy Module

Objective: Learn to manipulate structured datasets and perform numerical operations.

This module introduced the concept of structured data handling using the Pandas library. Students learned how to import datasets from various sources such as CSV, Excel, or SQL databases into DataFrames. Operations such as filtering, slicing, grouping, and merging were performed to manage and analyze large datasets.The NumPy library was taught for performing fast and efficient numerical operations. Participants learned about arrays, vectorized computations, and mathematical functions. Emphasis was placed on how NumPy arrays form the foundation for Pandas operations.By combining Pandas and NumPy, students learned how to transform raw data into a structured form that could be analyzed and visualized later.

Activities Undertaken:

- Creating Pandas DataFrames from lists/dictionaries.
- Indexing, slicing, filtering, grouping, and aggregating data.
- Handling arrays and performing numerical operations using NumPy.
- Using statistical functions: mean, median, standard deviation.

Learning Outcome:

- Students learned to perform numerical computations efficiently.
- Developed skills to manipulate large datasets using DataFrames and arrays.

2.4 Data Cleaning and Preprocessing

Objective: Prepare raw data for analysis.

This module addressed one of the most critical aspects of analytics—data cleaning. Real-world data often contains missing values, duplicates, inconsistencies, and errors that must be corrected before any analysis can be performed.Students learned methods to identify missing or null values using Pandas functions and strategies to handle them through imputation or deletion. They also learned how to remove duplicate entries to avoid redundancy. The concept of outliers was introduced, and boxplots were used to visualize data distribution and detect anomalies.Another important concept introduced was encoding categorical data—converting textual or categorical variables into numerical formats using methods like one-hot encoding, which is essential for data modeling.

Activities Undertaken:

- Identifying missing values and duplicates using df.isnull().sum() and df.duplicated().
- Handling missing data: median, mode, or dropping rows/columns.
- Detecting outliers using boxplots and statistical methods.
- Encoding categorical variables for analysis.

Figure 2.2 Steps of Data Cleaning

Learning Outcome:

- Gained understanding of ensuring data quality before analysis.
- Developed a systematic approach to cleaning and preprocessing datasets.

2.5 Data Visualization Module

Objective: Learn to represent data visually to identify patterns.

Visualization is one of the most powerful tools in data analytics as it helps in uncovering hidden patterns and trends. In this module, students explored Matplotlib and Seaborn, two widely used visualization libraries in Python. Participants created different types of charts, including bar plots for categorical data, histograms and boxplots for distributions, and scatter plots for relationship analysis. They also generated heatmaps to display correlations between variables .Trainers emphasized the selection of appropriate visualizations depending on the data type and analytical objective. Students were also introduced to customizing plots—titles, axes, legends, and color palettes—to make visualizations more readable and professional.

Activities Undertaken:

- Plotting categorical variables: bar plots, countplots.
- Visualizing numerical distributions: histograms, boxplots, scatter plots.
- Creating correlation heatmaps to understand relationships between features.

Learning Outcome:

- Developed the ability to interpret trends and patterns visually.
- Learned to select appropriate visualizations based on data type.

2.6 SQL and Database Management Module

Objective: Develop skills in querying and managing structured data.

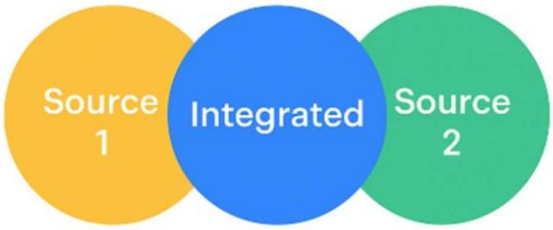
Data analytics often involves working with structured data stored in relational databases. This module introduced SQL (Structured Query Language) and MySQL, one of the most popular database management systems.

Students practiced basic SQL commands such as CREATE, SELECT, INSERT, UPDATE, and DELETE, collectively known as CRUD operations. They also learned to use filtering conditions (WHERE), sorting (ORDER BY), and grouping (GROUP BY) to extract specific insights.

The concept of JOINs was explained to demonstrate how data from multiple tables can be combined meaningfully. Towards the end, students learned how to connect Python with SQL databases, enabling end-to-end analytics where data is fetched from databases and analyzed using Python.

Activities Undertaken:

- Writing SQL queries: SELECT, WHERE, ORDER BY, GROUP BY.
- Aggregations and joins to combine multiple tables.
- Connecting Python with SQL databases for integrated analysis.



Learning Outcome:

- Ability to extract meaningful insights from structured databases.
- Learned integration of SQL and Python for end-to-end data handling.

2.7 Statistical Analysis Module

Objective: Apply statistical methods to datasets for insights.

This module aimed to provide students with a fundamental understanding of descriptive statistics, which are essential for any data analysis task. Concepts such as mean, median, and mode were discussed for measuring central tendency, while range, variance, and standard deviation were covered for measuring dispersion.Students also learned correlation analysis, which helps identify relationships between numeric variables. Through practical exercises, they were able to summarize datasets numerically and derive meaningful insights.The module emphasized how statistical understanding forms the foundation for advanced topics like predictive modeling and machine learning.

Activities Undertaken:

- Calculating measures of central tendency: mean, median, mode.
- Measures of dispersion: range, variance, standard deviation.
- Correlation analysis between numeric features.

Learning Outcome:

- Ability to summarize and interpret datasets numerically.
- Laid the groundwork for predictive modelling in future projects.

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Overview of Results

The 45-day industrial training provided an opportunity to apply theoretical knowledge of data analytics to practical scenarios. The training modules culminated in the project titled **Exploratory Data Analysis (EDA) on the Titanic Dataset**, which served as the capstone of the learning process. The objective of this project was to apply analytical and visualization techniques to a real-world dataset — the Titanic passenger data — in order to uncover meaningful insights, patterns, and relationships between variables that influenced survival.

The Titanic dataset is one of the most well-known datasets in data science, often used as a benchmark for learning data preprocessing, visualization, and predictive modeling. It provides demographic, travel, and fare-related details for over 800 passengers aboard the Titanic, along with their survival status.

The project involved several systematic steps, including data understanding, data cleaning, univariate and bivariate analysis, feature engineering, and interpretation of analytical results. Each of these stages is discussed in detail in the following sections.

3.2 Understanding the Dataset

The Titanic dataset was obtained from Kaggle's open data repository, which contains structured data related to the passengers on the ship RMS Titanic. The key variables in the dataset include:

Table 3.1 Features and Description of Titanic dataset

Feature	**Description**
PassengerId	Unique ID assigned to each passenger
Survived	Survival status (1 = Survived, 0 = Did not survive)
Pclass	Passenger class (1st, 2nd, or 3rd)
Name	Passenger's full name
Sex	Gender of the passenger
Age	Age of the passenger
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Ticket	Ticket number
Fare	Ticket fare
Cabin	Cabin number
Embarked	Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

3.2.1 Data Description

The dataset comprises a mix of numerical, categorical, and text-based attributes.

- **Numerical variables**: Age, Fare, SibSp, Parch
- **Categorical variables**: Sex, Pclass, Embarked

- **Mixed variables**: Ticket and Cabin (alphanumeric values)

Understanding these data types helped in determining the appropriate statistical and visualization techniques for analysis.

3.3 Data Preprocessing and Cleaning

Raw data typically contains inconsistencies, missing values, or noise that can affect analysis. Therefore, the dataset underwent a series of data cleaning and preprocessing operations to ensure analytical accuracy.

3.3.1 Handling Missing Values

- The *Age* column had several missing entries, which were filled using the median value to minimize bias.
- The *Embarked* column had a few missing records, replaced with the most frequent port (Southampton).
- The *Cabin* column contained numerous missing values, and since it wasn't crucial to the initial EDA, it was excluded from the main analysis.

3.3.2 Removing Redundant Columns

Columns such as Ticket, Name, and PassengerId were removed as they did not contribute directly to survival analysis.

3.3.3 Data Formatting

All categorical columns were standardized for consistent spelling and capitalization. Data types were converted into numeric or categorical forms wherever necessary for analysis.

3.4 Univariate Analysis

Univariate analysis involves examining each feature individually to understand its distribution, central tendency, and variability.

3.4.1 Numerical Features

- **Age**: Most passengers were between 20–40 years old. A smaller proportion of elderly passengers (>60 years) were observed.
- **Fare**: The fare distribution was highly skewed, with a few passengers paying exceptionally high fares.
- **SibSp and Parch**: Majority of passengers traveled alone or with one family member, highlighting a small family group trend.

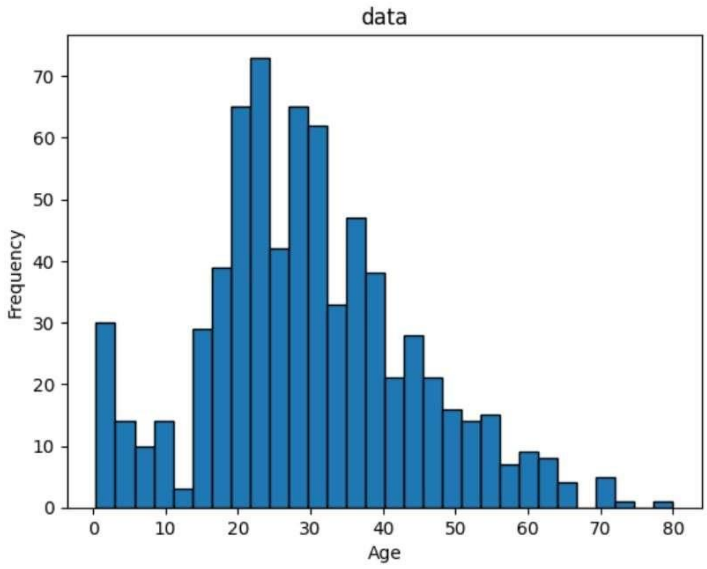


Figure 3.1 Frequency of passengers based on age

3.4.2 Categorical Features

- **Sex**: The dataset was male-dominated, with approximately 65% male passengers.
- **Pclass**: Third class had the maximum passengers, followed by first and second class.
- **Embarked**: Most passengers boarded from Southampton (S), followed by Cherbourg (C) and Queenstown (Q).

3.5 Bivariate Analysis

Bivariate analysis examines the relationship between two variables — crucial for discovering factors that influenced survival.

3.5.1 Categorical vs. Categorical

- **Gender vs. Survival:**

  A significant observation was that female passengers had a much higher survival rate than males. This supports the “women and children first” evacuation policy.

- **Pclass vs. Survival:**

  First-class passengers had the highest survival rate, while third-class passengers had the lowest. Economic status likely influenced access to lifeboats.

- **Embarked vs. Survival:**

  Passengers who boarded from Cherbourg showed slightly higher survival rates compared to those from Southampton.

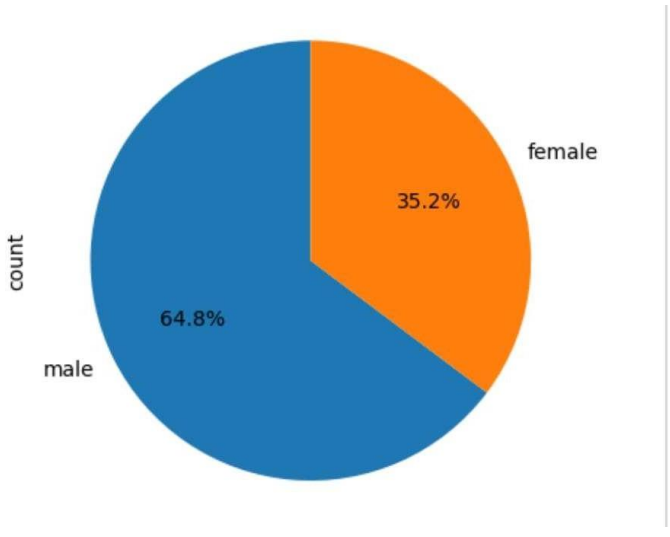


Figure 3.1 Distribution of died people based on gender

3.5.2 Numerical vs. Numerical

- **Age vs. Fare:**

  There was no strong correlation between age and fare, but most high fares corresponded to younger and middle-aged travelers from upper classes.

- **Fare vs. Survival:**

  A positive trend indicated that passengers who paid higher fares had higher survival chances.

3.5.3 Categorical vs. Numerical

- **Pclass vs. Fare:**

  A clear distinction was visible — average fare increased with higher class.

- **Survived vs. Age:**

  Children and younger adults had slightly better survival chances compared to older passengers.

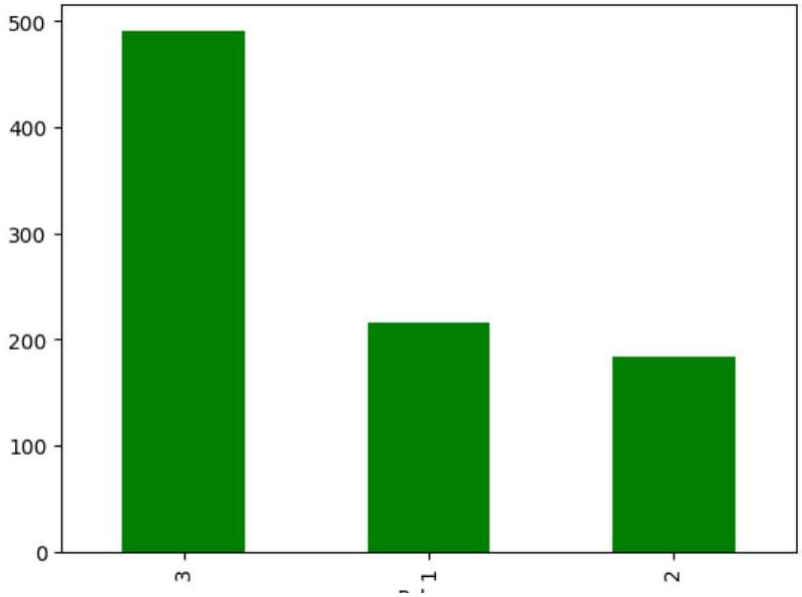


Figure 3.3 Distribution of passengers based on Pclass

3.6 Feature Engineering

To enhance the dataset and uncover deeper insights, new features were created from the existing data.

3.6.1 Family Size

A new variable Family_Size was derived as:

Family_Size = SibSp + Parch + 1

Passengers were categorized as:

- **Single**: Family size = 1
- **Small Family**: 2–4 members
- **Large Family**: 5 or more members

Observation:

Passengers with small families showed higher survival rates than singles or large families, likely due to easier coordination during evacuation.

3.6.2 Fare per Person

To normalize the fare, a new column Fare_Per_Person was derived as:

Fare_Per_Person = Fare / Family_Size

This revealed more realistic spending per individual, balancing the skewed fare distribution.

3.7 Correlation and Heatmap Analysis

A **correlation matrix** was computed to identify relationships between numerical features and the survival variable.

Key Observations:

- Positive correlation between Fare and Survived
- Negative correlation between Pclass and Survived (since higher class = lower numeric value)
- Weak correlation between Age and Survived

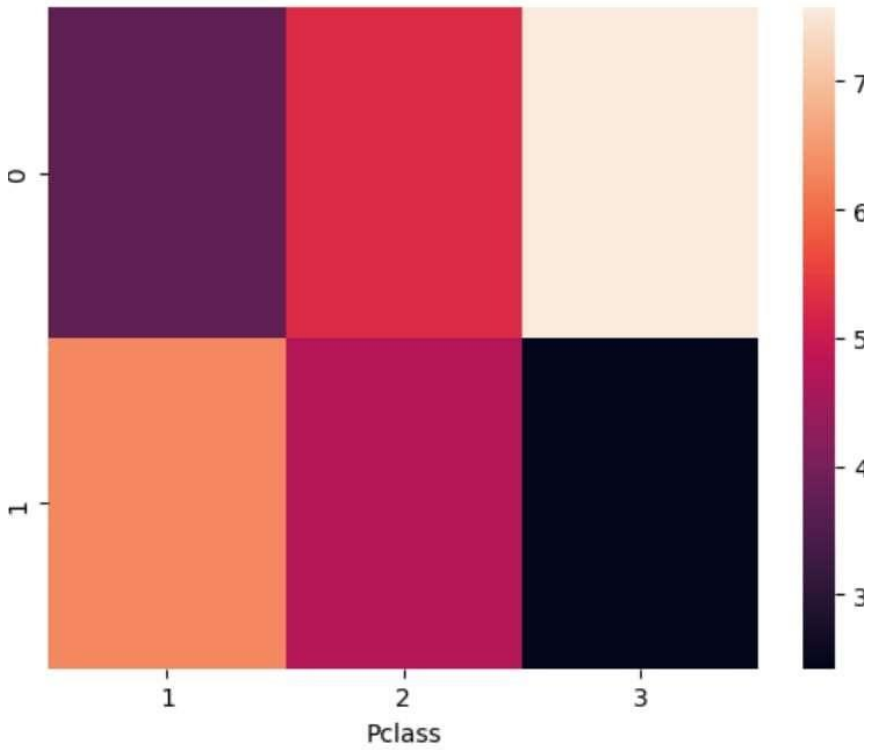


Figure 3.4 Heatmap Showing dead passengers based on Pclass

3.8 Visual Insights and Interpretations

Data visualization played a crucial role in interpreting findings clearly.

- **Class-wise Analysis:** First-class passengers enjoyed better survival prospects.
- **Age Distribution:** Children under 10 years had a moderately higher survival probability.
- **Fare Analysis:** Wealthier passengers tended to survive more, possibly due to better cabin locations and early rescue.

- **Survival Distribution by Gender:** Women survived in significantly higher proportions.

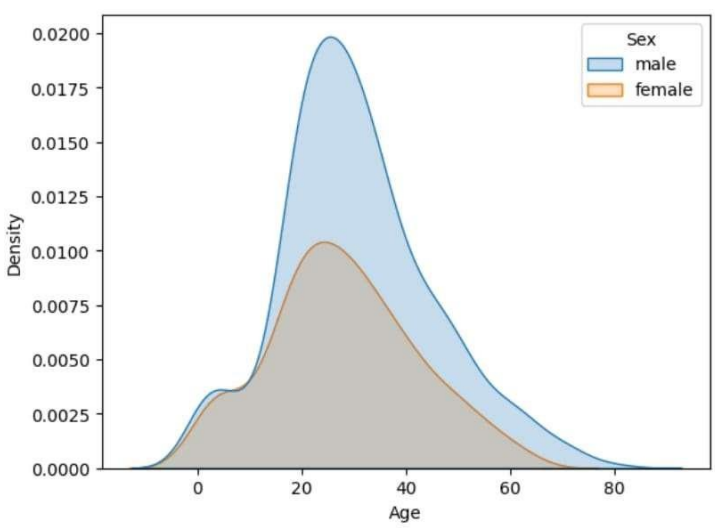


Figure 3.5 Survival Distribution Based On Gender

3.9 Discussion of Key Findings

The analytical study of the Titanic dataset revealed several meaningful insights into the social, demographic, and economic factors that influenced passenger survival. Through data visualization and statistical analysis, the project demonstrated how structured data can uncover real-world human patterns and validate historical accounts.

1. Socioeconomic Advantage

A strong correlation was observed between passenger class and survival rate. First-class passengers had a much higher chance of survival compared to those in second and third class. This highlights the impact of economic inequality, as upper-class passengers were located closer to lifeboats and received faster evacuation assistance. The analysis confirmed that wealth and cabin location played a crucial role in determining safety access during the disaster, emphasizing the real-world effects of social hierarchy in crisis situations.

2. Gender Influence

The data clearly showed that female passengers had a higher survival rate than males. This reflected the "Women and Children First" evacuation policy followed during the tragedy. Visualizations showed that over two-thirds of women survived, compared to only one-fifth of men. This pattern demonstrates how cultural norms and ethical decisions can strongly influence survival outcomes and how data analytics can help quantify such human factors.

3. Family Impact

Passengers traveling in small families (2–4 members) were found to have higher survival rates than those traveling alone or in large families. Smaller family groups likely provided mutual assistance and coordination, while large families may have faced confusion and delay during evacuation.
This insight, derived from feature engineering (using Family_Size = SibSp + Parch + 1), highlights how new variables can reveal hidden relationships and improve the interpretability of real-world datasets.

4. Age Factor

The analysis showed that younger passengers, especially children, had better survival chances. Older passengers faced difficulties due to reduced mobility and slower access to lifeboats. The Age vs. Survival visualization revealed that survival probability generally decreased with increasing age. Though not the strongest factor, age had a noticeable influence when combined with gender and class, offering deeper insight into demographic vulnerabilities.

5. Fare Relation

The fare amount paid by passengers was positively related to survival. Those who paid higher fares, usually first-class travelers, had better cabins and quicker access to lifeboats. Although fare alone did not determine survival, it indirectly represented economic privilege,

again linking wealth with safety. This finding supports the earlier observations about class-based survival differences.

Overall Interpretation

Overall, the analysis demonstrated that class, gender, family structure, age, and fare were the most influential factors affecting survival. The results not only support historical records but also show how data analytics can effectively model real-world human behavior. Through this project, it became clear that data science is a powerful tool for interpreting social patterns, inequalities, and decision-making processes during critical situations. The Titanic dataset thus served as a valuable medium for applying analytical techniques and understanding their relevance beyond technical boundaries.

CHAPTER 4

CONCLUSION AND FUTURE SCOPE

4.1 Conclusion

The 45-day industrial training was a significant step in bridging theoretical concepts with practical applications. Through this training, I gained hands-on exposure to data analysis, visualization, database querying, and project development.

The major outcomes of the training can be summarized as follows:

1. **Tool Proficiency:**
    - Learned to use Python, Pandas, NumPy, Seaborn, and Matplotlib for efficient data processing and visualization.
    - Gained practical knowledge of SQL for structured data querying and integration with Python.
2. **Project Development:**
    - Successfully completed an Exploratory Data Analysis (EDA) on the Titanic dataset, identifying key survival patterns with respect to gender, age, class, and fare.
    - Implemented data cleaning techniques, correlation analysis, and visualization models, which demonstrated the value of data-driven decision-making.
3. **Professional Skills Acquired:**
    - Improved logical thinking and problem-solving through real-world datasets.
    - Strengthened report writing and result interpretation skills.

- Understood the importance of data quality and preprocessing in analytics workflows.

In conclusion, the training not only enhanced my technical expertise but also helped in developing a structured approach to problem-solving and critical analysis, which will be beneficial for future professional endeavors.

4.2 Future Scope

While the training focused on foundational tools and a single dataset, there are numerous directions for extending this work.

1. **Advanced Machine Learning Models:**
    - Future projects can implement classification algorithms (Logistic Regression, Random Forest, XGBoost) to build predictive survival models.
    - Deep learning models could also be tested to improve accuracy.
2. **Big Data Integration:**
    - The skills learned can be extended to handle large-scale datasets using tools like Apache Spark and Hadoop.
    - Real-time data processing can be introduced for high-speed applications.
3. **Industry Applications:**
    - Similar analytics can be applied in healthcare (predicting disease risks), finance (fraud detection, customer segmentation), and transportation (traffic flow predictions).
4. **Visualization Dashboards:**
    - Development of interactive dashboards using Tableau, Power BI, or Plotly Dash would make data interpretation more user-friendly.

5. **Research Opportunities:**

    - The study can serve as a foundation for research in data ethics, fairness in AI, and socio-economic data analysis.

Thus, the project and training can be extended both academically and professionally to contribute toward solving real-world problems using data science.

REFERENCES

Books:

[1] W. McKinney, *Python for Data Analysis*, 2nd ed. Sebastopol, USA: O'Reilly Media, 2017.

[2] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Sebastopol, USA: O'Reilly Media, 2019.

Conference Papers:

[3] H. Chen, S. C. Laroiya, and M. Adithan, "Precision Machining of Advanced Ceramics," *International Conference on Advanced Manufacturing Technology (ICMAT-94)*, Johor Bahru, Malaysia, 1994, pp. 203–210.

Periodicals (Journals):

[4] R. E. Kalman, "New results in linear filtering and prediction theory," *Journal of Electrical Engineering*, vol. 83, no. 5, pp. 95–108, Mar. 1961.

[5] Y. V. Lavrova, "Geographic distribution of ionospheric disturbances in the F2 layer," *IET Microwaves, Antennas and Propagation*, vol. 19, no. 29, pp. 31–43, Feb. 1961.

Reports:

[6] E. E. Reber, "Oxygen absorption in the earth's atmosphere," *Aerospace Corporation*, Los Angeles, USA, Tech. Rep. TR-0200 (4230-46)-3, Nov. 1988.

Online Sources:

[7] R. J. Vidmar. (1994). *On the use of atmospheric plasmas as electromagnetic reflectors* [Online]. Available: ftp://atmnext.usc.edu/pub/etext/1994/atmosplasma.txt

[8] J. Jones. (1991, May 10). *Networks (2nd ed.)* [Online]. Available: http://www.atm.com

Thesis/Dissertations:

[9] J. O. Williams, "Narrow-band analyzer," Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, USA, 1993.

[10] N. Kawasaki, "Parametric study of thermal and chemical non-equilibrium nozzle flow," M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.

APPENDIX

Appendix A: Python Code for Titanic Dataset Analysis

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Load dataset
df = pd.read_csv("titanic.csv")

# Data cleaning
df['Age'].fillna(df['Age'].median(), inplace=True)
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
df.drop(columns=['Cabin', 'Ticket', 'PassengerId', 'Name'], inplace=True)

# Visualization 1: Survival by Gender
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title("Survival Distribution by Gender")
plt.show()

# Visualization 2: Survival by Class
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title("Survival Distribution by Passenger Class")
plt.show()

# Correlation Heatmap
sns.heatmap(df.corr(), annot=True, cmap="coolwarm")
plt.title("Correlation Heatmap of Titanic Dataset")
plt.show()
```

Figure 4.1 Python code for Titanic dataset